#2

500.40525X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): USHIJIMA, et al.

Serial No.: Not yet assigned

Filed: August 20, 2001

Title: INTEGRATED DATABASE SYSTEM AND PROGRAM STORAGE MEDIUM

Group: Not yet assigned

LETTER CLAIMING RIGHT OF PRIORITY

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231

August 20, 2001

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Patent Application No.(s) 2001-053474, filed February 28, 2001.

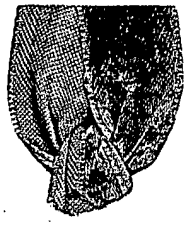
A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl I. Brundidge
Registration No. 29,621

CIB/alb
Attachment
(703)312-6600



日 本 国 特 許 庁
JAPAN PATENT OFFICE

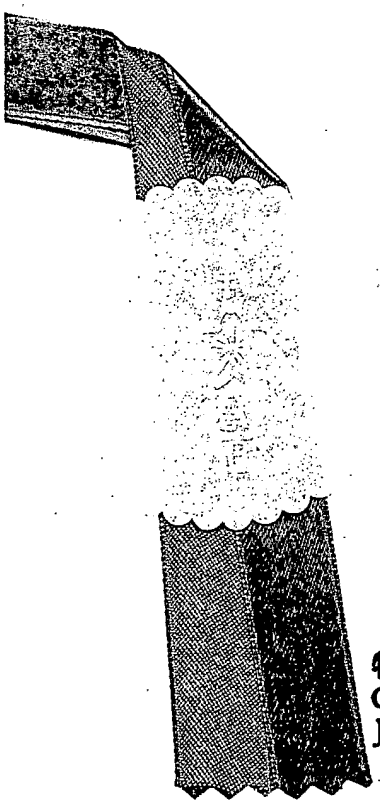
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００１年　２月２８日

出願番号
Application Number: 特願２００１－０５３４７４

出願人
Applicant(s): 株式会社日立製作所



CERTIFIED COPY OF
PRIORITY DOCUMENT

２００１年　７月２７日

特許庁長官
Commissioner,
Japan Patent Office

及川耕造

出証番号　出証特２００１－３０６５３７１

【書類名】　特許願
【整理番号】　H00014781A
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 17/30
【発明者】
【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所中央研究所内
【氏名】　牛嶋　一智
【発明者】
【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所中央研究所内
【氏名】　西澤　格
【発明者】
【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所中央研究所内
【氏名】　新谷　隆彦
【特許出願人】
【識別番号】　000005108
【氏名又は名称】　株式会社　日立製作所
【代理人】
【識別番号】　100075096
【弁理士】
【氏名又は名称】　作田　康夫
【電話番号】　03-3212-1111
【手数料の表示】
【予納台帳番号】　013088
【納付金額】　21,000円
【提出物件の目録】
【物件名】　明細書　1

出証特２００１－３０６５３７１

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　統合データベースシステムにおける問合せ最適化方法

【特許請求の範囲】

【請求項１】

複数の外部データベースを組合せて問合せ処理を行う統合データベースシステムにおいて、問合せで用いられる述語間の対応関係及び関連の強さに関する情報を含む述語辞書を備え、前記述語辞書を参照して、前記統合データベースシステムに対して投入された問合せを一つまたは複数の問合せ集合に変換する問合せ展開手段を含むことを特徴とする統合データベースシステム。

【請求項２】

請求項１記載の統合データベースシステムにおいて、問合せの変換を行うことが可能な述語間の対応関係が前記述語辞書に存在する場合に、前記述語辞書を参照した前記問合せ展開手段による問合せの変換を、繰返し適用することを特徴とする統合データベースシステム。

【請求項３】

請求項１記載の統合データベースシステムにおいて、さらに前記外部データベースの問合せ処理能力に関する仕様記述を備え、前記仕様記述を参照して、前記変換された問合せ集合の中から前記外部データベースを利用して実行可能な問合せを抽出する問い合わせ抽出手段を含むことを特徴とする統合データベースシステム。

【請求項４】

請求項１記載の統合データベースシステムにおいて、前記述語辞書に設定された関連の強さを参照して、前記変換された問合せ集合の中から適当な組合せの問合せを選択する問合せ選択手段を含むことを特徴とする統合データベースシステム。

【請求項５】

請求項１記載の統合データベースシステムにおいて、前記変換された問合せ集合のそれぞれの問合せを併合した問合せプランを生成する問合せプラン併合手段を含むことを特徴とする統合データベースシステム。

【請求項６】
複数の外部データベースを組合せて問合せ処理を行う統合データベースシステムを実現するプログラムを格納した記録媒体において、問合せで用いられる述語間の対応関係及び関連の強さに関する情報を含む述語辞書、及び前記述語辞書を参照して、前記統合データベースシステムに対して投入された問合せを一つまたは複数の問合せ集合に変換する問合せ展開処理を実現するプログラムを格納することを特徴とするプログラム記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】
本発明はデータベース、Ｗｅｂサーバ等のネットワークで接続された任意のデータソースを複数組合せて問合せ処理を行う統合データベースシステムに係わり、特に統合データベースに対して発行された問合せの最適化を行うシステムに関する。

【０００２】

【従来の技術】
ヒトの全ＤＮＡ配列データを読み取ることを目標として１９９０年から１５年を目処に開始されたヒトゲノム計画は、配列読み取り技術の大幅な進展並びに並列計算機を用いた大規模再配置手法の適用によりその読み取り速度が急速に加速され、２０００年６月には全ゲノムの約８５％に当たる部分について、そのドラフト配列の公表が行われるまでに至った。さらにヒトゲノム計画と並行して、ヒト以外の様々な生物種のＤＮＡ配列データについても解読が行われ、生体内のタンパク質のアミノ酸配列・立体構造データ、さらには代謝経路に関するデータなどについても解読結果がそれぞれの個別のデータベース（バイオ情報データベース）に整理・蓄積されている。これらのデータベースは、多くの場合何らかの公共機関によって管理運営され、インターネットを通じて参照することが可能となっている。

また、大量のＤＮＡ配列データの中から遺伝子をコードしている領域を予測したり、タンパク質の立体構造を予測するなど配列データから派生して得られる情報

を抽出する様々な解析ツールが次々と提案され、これらのツールの一部はインターネットを通じて公開されている。

一方、実験室における実験データ取得方法も大きく変革され、細胞内の多数の遺伝子の発現量を横断的に計測できるＤＮＡマイクロアレイ法など、高スループットで大量のデータを取得可能な手法が考案され、大量の実験結果が実験室内にも蓄積されるようになりつつある。今後は、これらデータベースやツールを組合せ、配列データにコードされている遺伝子やタンパク質が生体内においてどのような機能を担っており、どのように関連し合っているかを明らかにし、製薬・医療・食品等の各分野における応用を図っていく事が重要とされている。

生体現象の複雑な解析を行うためにはこれらのデータベースやツールを組合せた問合せ処理を行うことが不可欠であるが、その実現に際しては以下のような困難が伴う。

１．　収められたデータのデータ形式や問合せ形式が統一されておらず、複数のデータベースを組合せて利用する問合せを発行する事が困難である。

２．　収録されたデータの範囲や現象記述レベル等に関して、各データベースの問合せ能力がバラバラであるため、どのデータベースを関連させて問合せることが適当であるかを決定することが難しい。

３．　技術の進展に伴い、新たなデータベースやツールが次々と追加され、それぞれが個別に管理されているため、それらを既存のデータベース群と統合して利用するための手間が大きい。

今後、効率の良いバイオ情報解析を行うためには、これら複数のデータベースやツールを組合せた問合せを簡便に発行することが可能で、またこのような問合せを効率良く実行できるような統合データベースシステムを構築することが重要である。

複数のデータベースを組合せる統合データベースシステムにおいて効率の良い問合せ処理を行うためには、データ形式が異なる複数の外部データベースに対して、これらを関連付けた統合インターフェースを提供し、投入された問合せを効率良い問合せプランに変換して実行する問合せ最適化機構が重要である。

従来の統合データベースシステムにおける問合せ最適化方式としては、第１の方

式として、文献“ACM SIGMOD International Conference on Management of Data(SIGMOD'98)”(ACM Press発行)のP.564-566記載の論文 “Capability Based Mediation in TSIMMIS”及び米国特許第5588150号公報に開示されているラッパー・メディエータシステムにおけるアプローチ、第２の方式として、文献“Foundation of Intelligent Knowledge-Based Systems”(Academic Press発行)の12章“Multiagent systems”及び日本国公開特許平11-85522号に開示されているマルチエージェントシステムにおけるアプローチをあげることができる。

【０００３】

まず第１の従来方式であるラッパー・メディエータ型統合データベースシステムでは、個々の外部データベースに対してラッパーと呼ばれる問合せやデータ形式を変換するプログラムが用意され、メディエータは適切なラッパーを組合せることによって、複数のデータベースを単一のインターフェースを通じてアクセスすることを可能とする。このときそれぞれのラッパーは各自が受付可能な問合せのクラスを宣言しメディエータに登録する。投入された問合せの一部または全部がラッパーで宣言された問合せのクラスに含まれる場合は、その部分の処理をラッパー側に委託することが可能となる。メディエータは、ラッパー側の問合せ処理の見積りコスト等に基づき、その処理を委託するかどうかを決定する。

一般に、統合データベースシステムに対して投入された問合せを外部データベースを利用して処理する問合せプランは、利用する外部データベースの組合せおよび問合せ順序に関して何通りも考えることが出来き、これらの問合せプランは実行コスト及び問合せの結果得られるデータ内容に関してそれぞれ異なる性質を有する。しかし、第１の従来方式では、投入された問合せを外部データベースを利用して処理するいくつかの問合せプランの中から一つを選択して実行するため、得られる問合せ結果が本来外部データベースを利用して得られる問合せ結果より少なくなってしまう。

例えば、現在一般公開されているデータベースを用いてヒトゲノムに含まれる全ての遺伝子の集合を得ようとした場合、「遺伝子データベースに登録された遺伝子データからヒトのものであることが明示されているものを選ぶ」「配列データベースに登録されたヒトゲノムデータに対して遺伝子予測ツールを適用して抽出

する」「文献データベースに登録された文献から該当記述箇所を見つけ出し、そこで言及されている遺伝子名から決定する」など様々な問合せ方法を考えることが出来、それぞれの問合せ結果が持つ性質も大きく異なると予想される。そのため、第一の従来方式での問合せ最適化方式は、それぞれに格納されるデータや問合せ能力が互いに重なり合うようなデータベースが多数乱立しており、その組合せ方が幾通りもあるようなバイオ情報分野での統合データベース問合せ最適化方式としては適当でない。

【０００４】

次に第２の従来方式であるマルチエージェント型統合データベースシステムは、個々のデータソース及びデータソースに対する問合せ能力をカプセル化した外部エージェントと、投入された問合せを受け付け前記複数の外部エージェントにフォワードするコーディネートエージェントから成る。各外部エージェントはそれぞれが処理可能な問合せのクラスについて、予めコーディネートエージェントに対して登録しておき、ユーザから投入された問合せはコーディネートエージェントが各外部エージェントの登録内容に従って、問合せを処理可能な適当な外部エージェントに対して処理を委託する。このとき必要に応じてコーディネートエージェントは各外部エージェントで処理が可能なように問合せやデータ形式の変換を行う場合がある。

このように第２の従来方式では、ユーザに対して単一のインターフェースが提供されることはないが、個々のデータソースのデータ形式の違いや問合せ能力の違いをコーディネートエージェントが隠蔽することで、比較的簡便に問合せを発行することが出来る。

しかし第２の従来方式においても、エージェントの問合せ処理能力の包含関係に従って、投入された問合せの転送先のエージェントの組合せを一通りに決定して実行するため、第一の従来方式と同じく、得られる問合せ結果が本来外部データベースを利用して得られる問合せ結果より少なくなってしまう。そのため、バイオ情報分野での統合データベース問合せ最適化方式としては適当でない。

【０００５】

【発明が解決しようとする課題】

上述の従来方式では、投入された問合せを外部データベースでの問合せ処理を組み合わせて実行する幾通りの問合せプランの中から一つを選んで実行するため、この問合せプランを実行して得られる問合せ結果が本来外部データベースを利用して得られる問合せ結果より少なくなってしまう可能性が有る。

本発明の目的は、統合データベースシステムにおいて、ユーザが投入した問合せに対して外部データベースを組合せて利用する問合せプランを生成する際に、問合せ結果に求められる精度や問合せ処理のコストに対する要求を踏まえた上で、統合データベースにおける問合せ処理を効率良く行えるような問合せプランを生成する問合せ最適化方式を提供することである。

【０００６】

【課題を解決するための手段】

本発明の代表的な態様に従うシステムは、複数の外部データベースを組合せて問合せ処理を行う統合データベースシステムであって、問合せで用いられる述語間の関連を表す重み付きオントロジ、及び前記外部データベースの問合せ処理能力に関する仕様記述を備え、前記オントロジを参照して、前記統合データベースシステムに対して投入された問合せを一つまたは複数の問合せ集合に変換する問合せ展開手段、前記仕様記述を参照して前記変換された問合せ集合の中から前記外部データベースを利用して実行可能な問合せを抽出する問い合わせ抽出手段、前記オントロジに設定された重みを参照して、前記抽出された問合せ集合の中から適当な組合せの問合せを選択する問合せ選択手段、前記選択された問合せ集合のそれぞれの問合せを併合した問合せプランを生成する問合せプラン併合手段とを含む。

すなわち、投入された問合せに対して外部データベースを組合せて処理する問合せプランを複数生成し、生成された問合せプランの集合から投入された問合せに対してどれほど確からしい問合せ結果を返すことが出来るかを表す尺度としての近似度を参照して適切な問合せプランの部分集合を選択し、これに対して共通処理部分の括り出しなどの問合せ最適化を行いながら問合せプランを併合することで、問合せ処理コストを抑えつつ、投入された問合せに対してなるべく多くの問合せ結果を得ることの出来るような問合せプランを生成するようにする。

本発明の他の態様、およびこれらを実現するためより具体的なシステム構成については、実施例の説明において明らかにされる。

【０００７】

【発明の実施の形態】

図１に本発明における問合せ最適化方式を備える統合データベースシステムの実施形態の一例を示す

本実施例における統合データベースシステム１は、ユーザまたはプログラムから発行された問合せを受け付け、問合せ投入時あるいは投入前に指定された問合せ結果の精度や問合せコストに関する指定を参照しながら、この問合せに対して予め指定された外部データベース群２を組合せる事で処理内容を近似するような近似問合せを複数生成し、これら近似問合せの問合せプランを併合することで、投入された問合せに対してより効率的な近似を行う問合せプランを生成する。統合データベースシステムは、この問合せプランの実行結果を最初に投入された問合せの近似された問合せ結果として返す。

図１において、オントロジモジュール３は前記統合データベースシステム１が統合を行う前記外部データベース２との間の関係を表すオントロジを保持する。ディクショナリモジュール４は前記外部データベース２に関する仕様記述を保持する。問合せ受付モジュール５はユーザまたはプログラムから発行された問合せを受け付け、統合データベースシステムでの内部表現への変換を行う。問合せ最適化モジュール６は、前記問合せ受付モジュール５が生成した問合せの内部表現及び前記オントロジを参照して、問合せの内部表現を変換しながら前記外部データベースを組合せて利用する近似問合せの集合を生成する。問合せプラン生成モジュール７は、前記問合せ最適化モジュールが生成した近似問合せ集合のそれぞれの近似問合せに対して問合せプランを生成し、これらを併合することで最終的な問合せプランを生成する。

以下に、上記各構成要素の詳細構造について述べる。

「オントロジモジュール」

本実施例においてオントロジモジュール３が保持するオントロジ１４は、図２に示すような有向グラフで表される。オントロジは統合データベースシステムに対

する問合せで利用される述語や外部データベースを整理分類するグラフ構造として利用される。オントロジのノード２１及び有向エッジ２２はそれぞれ問合せの内部表現として利用される述語及び述語の間の対応関係を表す。また統合データベースに対して発行された問合せは、問合せ受付モジュール５においてオントロジ上の述語を用いた内部表現に変換される。

本実施例におけるオントロジでは、ノードＮに対応する述語が他のノード群Ｎｉに対応する述語の組合せで近似可能な場合、ノードＮからノード群Ｎｉに対してエッジが張られる。このときエッジに対しては、ノードＮに対応する述語をノード群Ｎｉに対応する述語で置き換える場合の対応関係を表す“近似ルール”と近似ルールで対応付けられた述語間の関連の度合いを表す“近似度”が設定される。

「外部データベース群」

本実施例における外部データベース２は、問合せ対象となるデータソース１８及び前記データソースに対して一定の問合せ処理を行なう検索ツール１７の組合せから成る。

それぞれの外部データベースのデータソース１８にどのようなデータが蓄積されており、また外部データソースの検索ツール１７がどのような問合せ処理を行えることができるかに関する外部データベース仕様記述１５は、前記ディクショナリモジュール４に格納される。

「ディクショナリモジュール」

本実施例におけるディクショナリモジュール４は、前記外部データベース２に関する外部データベース仕様記述１５に加え、統合データベースシステムが変換することのできるデータ型や問合せのクラスの間の対応付けに関するデータ・問合せ変換仕様記述１６を保持する。

「問合せ受付モジュール」

本実施例における問合せ受付モジュール５は、統合データベースシステム１に対して投入された問合せを前記オントロジ１４上の述語を用いた内部表現に変換する問合せ変換手段８を保持する。統合データベースシステム１に対して発行される問合せに対しては、問合せ投入時あるいは投入前に近似問合せを生成する際の

近似度の下限及び実行コストの上限を指定することができる。

「問合せ最適化モジュール」

本実施例における問合せ最適化モジュール６は、投入された問合せを近似問合せ候補の集合に展開する問合せ展開手段９、展開された近似問合せ候補の集合から、ディクショナリモジュールに登録された外部データベース等の仕様記述を参照して、実行可能な問合せを抽出する実行可能問合せ抽出手段１０、抽出された実行可能な近似問合せの集合から、指定された近似度の下限と実行コストの上限に照らして最適な近似問合せ候補の集合を選択する最適近似問合せ選択手段１１から成る。

「問合せプラン生成モジュール」

本実施例における問合せプラン生成モジュール７は、前記問合せ最適モジュールで選択された最適近似問合せ集合のそれぞれの近似問合せに対して問合せプランを生成する問合せプラン生成手段１２、生成された問合せプランの集合から投入された問合せに対して近似度の大きさなど適当な順番でプランの併合を行い、近似度と実行コストに関して最適な問合せプランの組合せを生成する問合せプラン併合手段１３から成る。

「全体のフローチャートとその説明」

次に図３において本実施例における処理全体の処理フローの様子を示す。

まず、ユーザまたはプログラムによって統合データベースシステム１に対して発行された問合せは、前記問合せ受付モジュール５の問合せ変換手段８によって前記オントロジ１４上の述語を用いた内部表現に変換される（ステップ３１）。このとき、投入された問合せを近似問合せに展開する際の近似度の下限と問合せプランを作成する際の実行コストの上限を予め指定しておくか、問合せ投入時に指定することができる（ステップ３０）。

続いて内部表現に変換された問合せは、前記問合せ最適化モジュール６の近似問合せ展開手段９によって、オントロジ１４上のノードに設定された近似ルールに従って、指定された近似度の下限を下回らない範囲で近似問合せの集合へ展開される（ステップ３２）。

ついで前記問合せ最適化モジュール６の実行可能問合せ抽出手段１０によって、

近似問合せ集合のうち実行可能な問合せプランが生成できる近似問合せのみが抽出され近似問合せ集合に残される。さらに前記問合せ最適化モジュール６の最適近似問合せ選択手段１１によって、近似問合せ集合内の近似問合せの内、投入された問合せを一定の基準において最も良く近似すると思われる組合せが選択され、これが近似問合せ集合へ残される（ステップ３３）。近似問合せ集合内の近似問合せは全てその近似値の降順に関して整列される（ステップ３４）。その後、前記問合せプラン生成モジュール７の問合せプラン生成手段１２にて問合せプランに変更されながら、指定された実行コストを超えないように順に併合され（ステップ３５）、最終問合せプランが得られた時点でこれが実行される（ステップ３６）。

「各モジュールの手段毎のフローチャートとその説明」

以下では、各モジュールを構成する手段毎の詳細処理フローの様子を示す。

一般に本実施例における近似問合せは、{(n1,..,ni)，a0}と表現される。ここでniはオントロジ上の述語、aは近似度を表し、最初に投入された問合せの場合は必ずa=1.0である。このとき、この一般形で表される近似問合せでは、統合データベースシステムに対して述語n1,..,述語niによって指定される条件を満たすようなデータの集合を問合せ結果として返すように指示しており、近似度a0はその場合の問合せ結果が、元の問合せの問合せ結果とおおよそa0の割合で一致すると予想されることを示している。

図４及び図５に、前記問合せ最適化モジュール６の近似問合せ展開手段９の処理フローを示す。

本実施例における近似問合せ展開においては、まず前記問合せ受付モジュール５によってオントロジ上の述語を利用して変換された内部表現{(n1,..,ni)，1.0}が、近似問合せ集合ＳＱの最初の要素として設定される（ステップ４０）。続いて、近似問合せ集合ＳＱ内のそれぞれの近似問合せｓｑ＝｛(n1,..,ni)，a0｝について、述語niに対応するオントロジ上のノードＮｉに対して、以下のような問合せ近似ルール及び近似度が設定されている場合、

｛ni→（m1,...，mm），a1｝

元の近似問合せｓｑの述語niをm1,...,mmで置き換え、次式の近似問い合わせ、

つまり近似度をa0＊a1とした新しい近似問合せｎｓｑを生成し（図５のステップ５１）、

ｎｓｑ＝｛（n1,...,ni-1, m1,...,mm,ni+1,...,nn），a0＊a1｝

これを近似と合い合わせ集合ＳＱに加える（図４のステップ４２）。

この操作は、近似問合せ集合ＳＱにオントロジ上の近似ルールによって書き換え可能な近似問合せが存在するまで繰り返される(４３)。ただし、展開された結果の近似問合せの近似度が指定された下限値を下回る場合は、その近似問合せを近似問合せ集合ＳＱから削除してしまっても良い。

図６は、前記問合せ最適化モジュールにおける実行可能問合せ抽出手段１０と最適近似問合せ選択手段１１の処理フローを示す。

まず実行可能問合せ抽出手段１０では、近似問合せ集合ＳＱ内の全ての近似問合せｓｑ＝｛（n1,...,ni,...,nn），ａ0｝について、近似問合せの問合せ部に出現する全ての述語niについてオントロジを参照し、述語niに対応するノードNiに利用可能な要素データベースが存在し、かつ実行可能な問合せプランを構成可能なものだけを抽出し、近似問合せ集合ＳＱに残す（ステップ６０）。

ついで、最適近似問合せ選択手段１１では、近似問合せ集合ＳＱ内に残された問合せｓｑ＝｛（n1,...,ni,...,nn），ａ0｝の中から元の問合せの問合せ結果を効率良く近似するような適当な近似問合せの組合せを選択し、これらを近似問合せ集合ＳＱに残す（ステップ６１）。近似問合せの組合せの選択方法は、ここでは特に規定しない。

図７は、前記問合せプラン生成モジュールにおける問合せプラン生成手段１２と問合せプラン併合手段１３の処理フローを示す。

まずはじめに、問合せプラン生成手段によって前記近似問合せ集合内の近似問合せを問合せプランに変換し（ステップ７０）、ついで変換された問合せプランを近似度の大きい順に取り出しながら共通処理部分の括り出しを行うなどして併合する（ステップ７２）。これを指定された問合せコストの上限を超えないように繰返し（ステップ７３）、最終的に得られた問合せプランを最終プランとし、これを実行する（ステップ７５）。

但し、上記で示した最適化方法は実施形態の一例であり、本発明はこれに限定さ

れるものではない。

「具体的な処理に対する適用例の説明」

以下では、本実施例における問合せ最適化処理を具体的な問合せ例に適用した場合について説明する。

まず、統合データベースシステムで利用される外部データベースとして図８に示すような４つの外部データベース群を想定する。すなわち、ＤＢ１は、これまでに解読されたタンパク質のアミノ酸配列・立体構造や機能についての情報などを蓄積するタンパク質ＤＢである。ＤＢ２は、これまでに知られている様々な酵素反応の反応式を蓄積する酵素反応ＤＢである。ＤＢ３は、これまで知られている遺伝子の転写調節因子に関する情報を蓄積する転写調節因子ＤＢである。ＤＢ４は様々な生物種のゲノム配列情報を蓄積するゲノム配列ＤＢである。

【０００８】

また、このとき統合データベースシステムで利用されるオントロジとして、図８に示す有向グラフ構造を利用する。このオントロジは、７種類の述語を含み、それらの述語間の近似変換ルールとして８種類のルールが設定されている。

すなわち述語は７種類であり、列挙すると以下の通りである。、

発現抑制（遺伝子：gX，遺伝子：gY）：遺伝子gXが遺伝子gYの発現量を抑制する関係にあることを示す。

機能障害（タンパク質：pX，タンパク質：pY）：タンパク質pXがタンパク質pYに作用してタンパク質pYの機能を障害する関係にあることを示す。

機能促進（タンパク質：pX， タンパク質：pY）：タンパク質pXがタンパク質pYに作用してタンパク質pYの機能を促進する関係にあることを示す。

タンパク質機能（種別：X，タンパク質：pX，タンパク質：pY）：タンパク質pXがタンパク質pYに対して種別X（＝機能失活等）の機能を有することを示す。

酵素反応（タンパク質：pE，タンパク質：pX，タンパク質：pY）：タンパク質pEがタンパク質pXからタンパク質pYへの反応を触媒する酵素であることを示す。

転写調節因子（種別：X，タンパク質：pX，遺伝子：gY）：タンパク質pXは遺伝子gYの種別X（＝エンハンサ、リプレッサ等）の転写調節因子であることを示す。

配列（種別：X，遺伝子：gX，タンパク質：pX）：遺伝子gXがタンパク質pXに対して種別X（＝エンコード等）の関係にあることを示す。

近似ルールは、以下の８種類が設定されているとする。

Ｒ１：エンハンサ機能阻害＝｛発現抑制（遺伝子：gX，遺伝子：gY）→（配列（種別：エンコード，遺伝子：gX，タンパク質：pX），機能阻害（タンパク質：pX，タンパク質：enY），転写調節因子（種別：エンハンサ，タンパク質：enY，遺伝子：gY）），０．５｝

Ｒ２：リプレッサ機能促進＝｛発現抑制（遺伝子：gX，遺伝子：gY）→（配列（種別：エンコード，遺伝子：gX，タンパク質：pX），機能促進（タンパク質：pX，タンパク質：reY），転写調節因子（種別：リプレッサ，タンパク質：reY，遺伝子：gY）），０．５｝

Ｒ３：機能失活＝｛機能阻害（タンパク質：pX，タンパク質：pY）→（タンパク質機能（種別：失活，タンパク質：pX，タンパク質：pY）），０．７｝

Ｒ４：生成酵素失活＝｛機能阻害（タンパク質：pX，タンパク質：pY）→（タンパク質機能（種別：失活，タンパク質：pX，タンパク質：pE），酵素反応（タンパク質：pE，タンパク質：pre-pY，タンパク質：pY），０．２｝

Ｒ５：転写因子抑制＝｛機能阻害（タンパク質：pX，タンパク質：pY）→（転写調節因子（種別：リプレッサ，タンパク質：pX，遺伝子：g-enY），配列（種別：エンコード，遺伝子：g-enY，タンパク質：enY）），０．１｝

Ｒ６：前駆体＝｛機能促進（タンパク質：pX，タンパク質：pY）→（酵素反応（タンパク質：pE，タンパク質：pX，タンパク質：pY）），０．３｝

Ｒ７：生成酵素＝｛機能促進（タンパク質：pX，タンパク質：pY）→（酵素反応（タンパク質：pX，タンパク質：pre-pY，タンパク質：pY）），０．５｝

Ｒ８：転写因子促進＝｛機能促進（タンパク質：pX，タンパク質：pY）→（転写調節因子（種別：エンハンサ，タンパク質：pX，遺伝子：g-pY），配列（種別：エンコード，遺伝子：g-pY，タンパク質：pY）），０．２｝

ここで図９は、それぞれ近似ルールＲ３からＲ８における各因子の関連を模式的に示す。

前述の４つのＤＢはそれぞれ、オントロジ上のいくつかの述語に関する問合せを

処理する能力を有し、それぞれの述語に対応するオントロジ上のノードと対応付けられている。すなわち
タンパク質ＤＢは、述語“タンパク質機能”に関する問合せを受け付けることが可能、酵素反応ＤＢは、述語“酵素反応”に関する問合せを受け付けることが可能、転写調節因子ＤＢは、述語“転写調節因子”に関する問合せを受け付けることが可能、ゲノム配列ＤＢは、述語“配列”に関する問合せを受け付けることが可能とする。

【０００９】

このとき例えば、ある生物の細胞への新規遺伝子ｇＸの導入実験に先立って、統合データベースに対して遺伝子gXの導入によって発現量が減少することが予測される遺伝子gＹを求める、Ｑ＝発現抑制（遺伝子：gX，遺伝子：gY）という問合せが、近似度の下限MinApprox＝０．１、及び実行コストの上限MaxExecCost＝１０００という指定と共に発行されたとする。
前記問合せを受け付けた問合せ受付モジュールは、投入された問合せを対応する述語を表すノードＮ０に対応付け、
Ｓ＝[Ｐ０＝｛Ｎ０：発現抑制（遺伝子：gX，遺伝子：gY），１．０｝]
なる内部表現に変換する。
続いて問合せ最適化モジュールはノードＮ０に接続しているエッジＥ１，Ｅ２を参照し、そこに対応付けられている近似ルールＲ１，Ｒ２をそれぞれ用いて投入された問合せを変形し、新たな近似問合せＰ１、Ｐ２を近似問合せ集合Ｓに追加する。
第一回目適用：
Ｓ＝[
Ｐ０＝｛Ｎ０：発現抑制（遺伝子：gX，遺伝子：gY），１．０｝，
Ｐ１＝｛Ｎ６：配列（しゅべつ：エンコード，遺伝子：gX，タンパク質：pX），
　　Ｎ１：機能障害（タンパク質：pX，タンパク質：enY），
　　Ｎ５：転写調節因子（種別：エンハンサ，タンパク質：enY，遺伝子：gY），０．５｝，
Ｐ２＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ２：機能促進（タンパク質：pX，タンパク質：pre-reY，タンパク質：reY），

Ｎ５：転写調節因子（種別：リプレッサ，タンパク質：reY，遺伝子：gY），０．５｝］

問合せ最適化モジュールは、この操作を新たに問合せが対応付けられたノードについて繰返し適用してゆくことで、投入された問合せを近似問合せの集合へ展開する。このときノードに適用される問合せの近似度が近似度の下限MinApproxを下回った場合は、その問合せは近似問合せの集合から削除され、以後考慮の対象外となる。また、問合せに出現する述語に対して適用可能な近似ルールがなくなった時点で、投入された問合せの近似問合せの集合への展開処理は終了する。

第二回目適用：

Ｓ＝［

Ｐ０＝｛Ｎ０：発現抑制（遺伝子：gX，遺伝子：gY），１．０｝，

Ｐ１＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ１：機能障害（タンパク質：pX，タンパク質：enY），

Ｎ５：転写調節因子（種別：エンハンサ，タンパク質：enY，遺伝子：gY），０．５｝，

Ｐ２＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ２：機能促進（タンパク質：pX，タンパク質：pre-reY，タンパク質：reY），

Ｎ５：転写調節因子（種別：リプレッサ，タンパク質：reY，遺伝子：gY），０．５｝，

Ｐ３＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ３：タンパク質機能（関係：失活，タンパク質：pX，タンパク質：enY），

Ｎ５：転写調節因子（種別：エンハンサ，タンパク質：enY，遺伝子：gY），０．１５｝，

Ｐ４＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ３：タンパク質機能（関係：失活,タンパク質：pX，タンパク質：pE

），

Ｎ４：酵素反応（タンパク質：pE，タンパク質：pre-enY，タンパク質：enY），

Ｎ５：転写調節因子（種別：エンハンサ，タンパク質：enY，遺伝子：gY），０．０５｝，

Ｐ５＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ５：転写調節因子（種別：リプレッサ，タンパク質：pX，タンパク質：g-enY），

Ｎ６：配列（種別：エンコード，遺伝子：g-enY，タンパク質：enY），

Ｎ５：転写調節因子（種別：エンハンサ，タンパク質：enY，遺伝子：gY），０．０５｝，

Ｐ６＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ４：酵素反応（タンパク質：pE，タンパク質：pX，タンパク質：reY），

Ｎ５：転写調節因子（種別：リプレッサ，タンパク質：reY，遺伝子：gY），０．１５｝，

Ｐ７＝｛Ｎ６：配列（種別：エンコード，遺伝子：gX，タンパク質：pX），

Ｎ４：酵素反応（タンパク質：pX，タンパク質：pre-reY，タンパク質：reY），

Ｎ５：転写調節因子（種別：リプレッサ，タンパク質：reY，遺伝子：gY），０．２５｝，

Ｐ８＝｛Ｎ６：配列（種別：エンコード（遺伝子：gX，タンパク質：pX），

Ｎ５：転写調節因子（種別：エンハンサ，タンパク質：pX，遺伝子：g-reY），

Ｎ６：配列（種別：エンコード，遺伝子：g-reY，タンパク質：reY），

Ｎ５：転写調節因子（種別；リプレッサ，タンパク質：reY，遺伝子：gY），０．１｝］

このとき、近似度の下限が０．０１であれば、上記の全ての近似問合せが近似問合せの集合Ｓに残されるが、この例で指定された近似度の下限は０．１であるの

で、上記のうち６番目の近似問合せＰ５は利用されない。また問合せプラン生成モジュールでは、前記問合せ最適化モジュールで決定された近似問合せの集合の中から実行可能な近似問合せの抽出を行う。すなわちある近似問合せが実行可能であるためには、以下の二点が必要である。

１．近似問合せを構成するそれぞれの述語に対して利用可能な外部データベースが存在すること。

２．外部データベースや検索ツールに対して適用される問合せが実行可能となるような問合せプランを生成であること。

すなわち例えば、上述の例において外部データベースのタンパク質ＤＢ（ＤＢ１）が利用可能でない場合は、ノードＮ３で利用可能な外部データベースが存在しないため、上記の近似問合せのうちＰ０からＰ４は実行可能ではない。従って上記近似問合せの内、実行可能なものはＰ６～Ｐ８である。

続いて問合せプラン生成モジュールは、実行可能なＰ６からＰ８の近似問合せに対して、各外部データベースが受理可能な問合せプランを生成する。例えば近似問合せＰ６の場合、遺伝子gXを指定してゲノム配列ＤＢに問合せ、その結果得られたタンパク質pXの値で酵素反応ＤＢを問合せ、さらにその結果得られたタンパク質reYの値で転写調節因子ＤＢを問合せる図１０に示すような問合せプランを考えることが出来る。

続いて問合せプラン生成モジュールは、生成された問合せプランを近似度の大きい順に併合することを試みる。この例の場合ゲノム配列ＤＢに対する問合せ｛Ｎ６：エンコード（遺伝子：gX，タンパク質：pX）｝と転写調節因子ＤＢに対する問合せ｛Ｎ５：リプレッサ（タンパク質：reY，遺伝子：gY）｝は共通であるのでこれを併合することで、図１１に示すような問合せプランが生成される。この問合せプランでは、遺伝子gXを指定してゲノム配列ＤＢに問合せ、その結果得られたタンパク質pXの値で酵素反応ＤＢおよび転写調節因子ＤＢを問合せ、転写調節因子ＤＢを問合せた結果得られた遺伝子g-reYの値で再びゲノム配列ＤＢを問合せ、最後に酵素反応ＤＢを問合せた結果と最後にゲノム配列ＤＢを問合せた結果得られたタンパク質reYの値で転写調節因子ＤＢを問合せている。ただし、このとき併合後の問合せプランのコストは指定された問合せコストの上限を超えなか

ったものとする。

【００１０】

【発明の効果】

本発明による問合せ最適化では、投入された問合せを構成する述語を、述語間の関連度に関する重みが設定されたオントロジを利用して書き換えることで、当該問合せを複数の近似問合せに展開し、これらのいくつかを纏めて実行することにより、より多くの問合せ結果を効率的に取得することが可能となる。

【図面の簡単な説明】

【図１】

本発明の実施例の統合データベースシステムを示すブロック図である。

【図２】

上記実施例のオントロジの模式図である。

【図３】

上記実施例の問合せ最適化のフローチャートである。

【図４】

上記実施例の近似問合せ展開手段のフローチャートである。

【図５】

近似問合せ展開手段のフローチャートである。

【図６】

実行可能問合せ抽出手段及び最適近似問合せ選択手段のフローチャートである。

【図７】

問合せプラン生成手段及び問合せプラン併合手段のフローチャートである。

【図８】

実施例におけるオントロジの一例である。

【図９】

実施例における近似ルールに対応する生体分子の関連模式図である。

【図１０】

実施例における問合せプランの一例である。

【図１１】

実施例における併合された問合せプランの一例である。

【符号の説明】

１：統合データベースシステム

２：外部データベース群

３：オントロジモジュール

４：ディクショナリモジュール

５：問合せ受付モジュール

６：問合せ最適化モジュール

７：問合せプラン生成モジュール

８：問合せ変換手段

９：近似問合せ展開手段

１０：実行可能問合せ抽出手段

１１：最適近似問合せ選択手段

１２：問合せプラン生成手段

１３：問合せプラン併合手段

１４：オントロジ

１５：外部データベース仕様記述

１６：データ・問合せ変換仕様記述

１７：検索ツール

１８：データソース

２０：問合せ述語

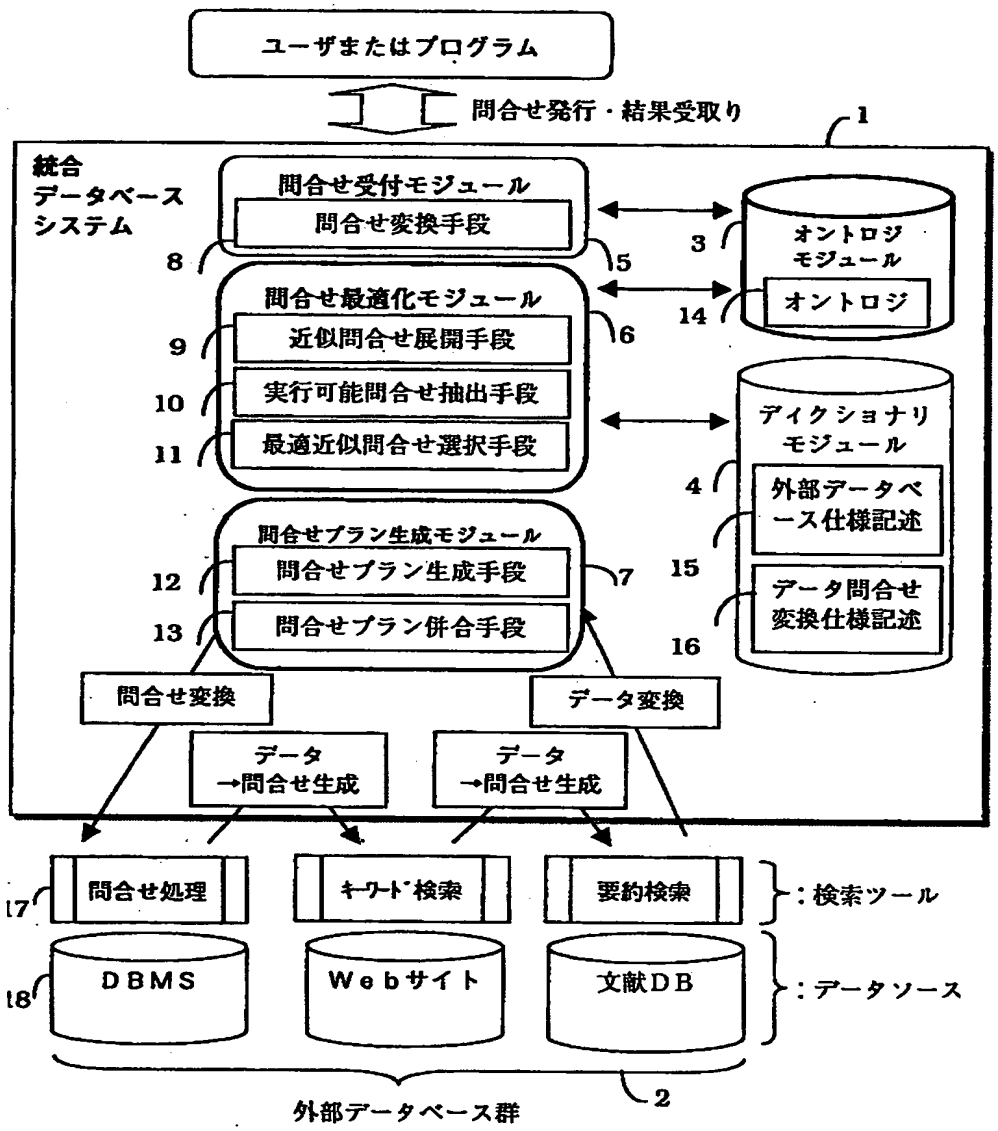
２１：ノード

２２：エッジ。

【書類名】　図面

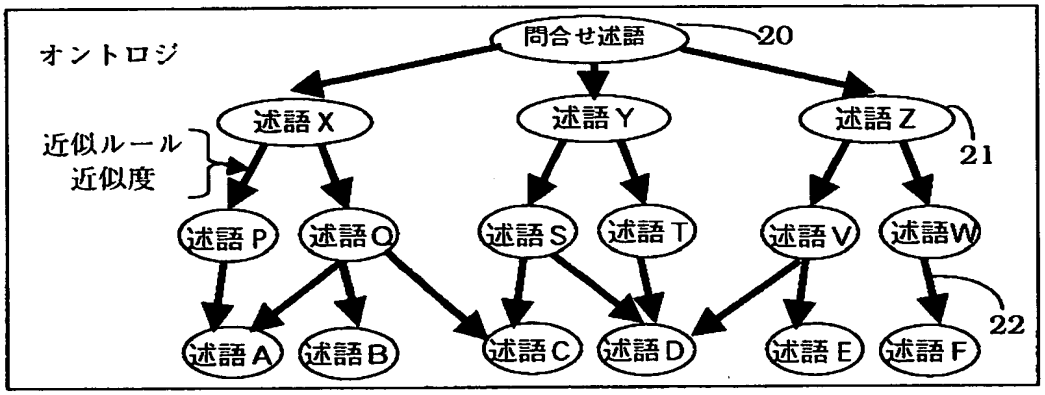
【図１】

図1


ユーザまたはプログラム
問合せ発行・結果受取り
1
統合
データベース
システム
問合せ受付モジュール
問合せ変換手段
8
5
3
オントロジ
モジュール
オントロジ
14
問合せ最適化モジュール
9
近似問合せ展開手段
6
10
実行可能問合せ抽出手段
11
最適近似問合せ選択手段
ディクショナリ
モジュール
4
外部データベース仕様記述
問合せプラン生成モジュール
12
問合せプラン生成手段
7
15
データ問合せ変換仕様記述
13
問合せプラン併合手段
16
問合せ変換
データ変換
データ
→問合せ生成
データ
→問合せ生成
17
問合せ処理
キーワード検索
要約検索
：検索ツール
18
DBMS
Webサイト
文献DB
：データソース
2
外部データベース群

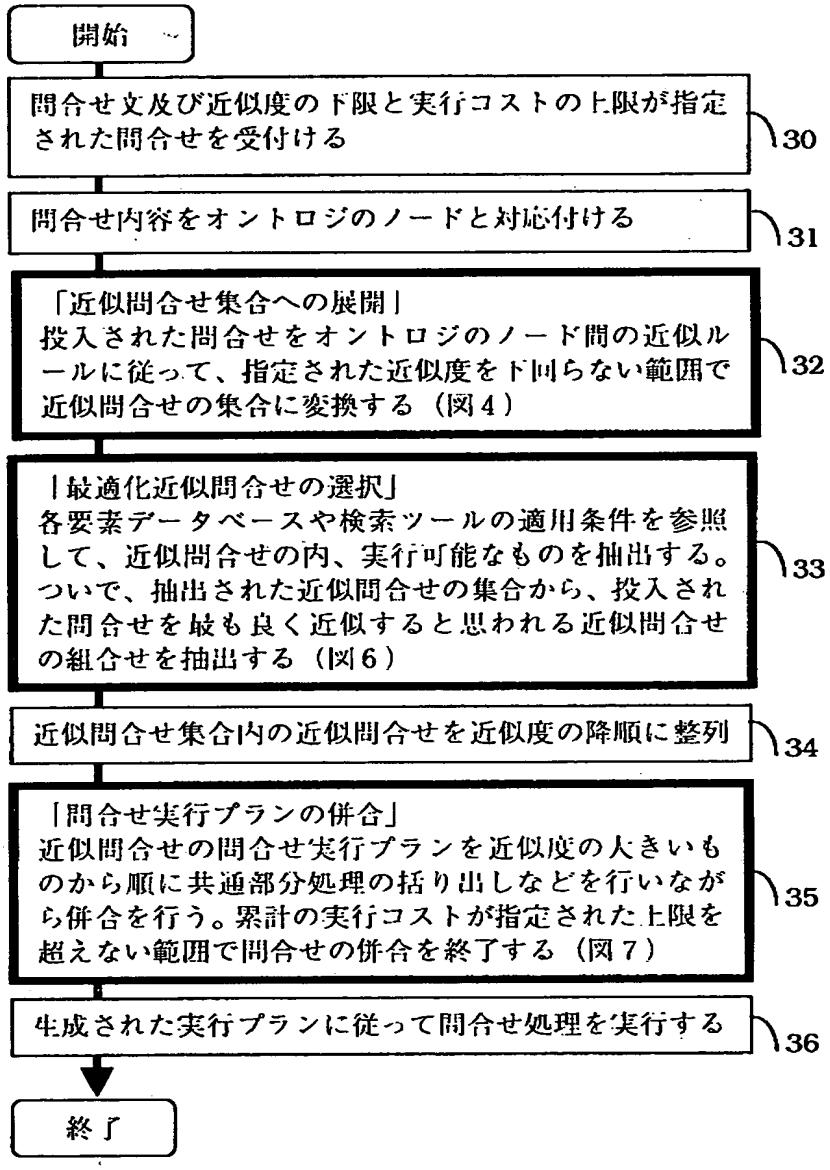
【図２】

図２


オントロジ
問合せ述語
20
述語Ｘ
述語Ｙ
述語Ｚ
21
近似ルール
近似度
述語Ｐ
述語Ｑ
述語Ｓ
述語Ｔ
述語Ｖ
述語Ｗ
22
述語Ａ
述語Ｂ
述語Ｃ
述語Ｄ
述語Ｅ
述語Ｆ

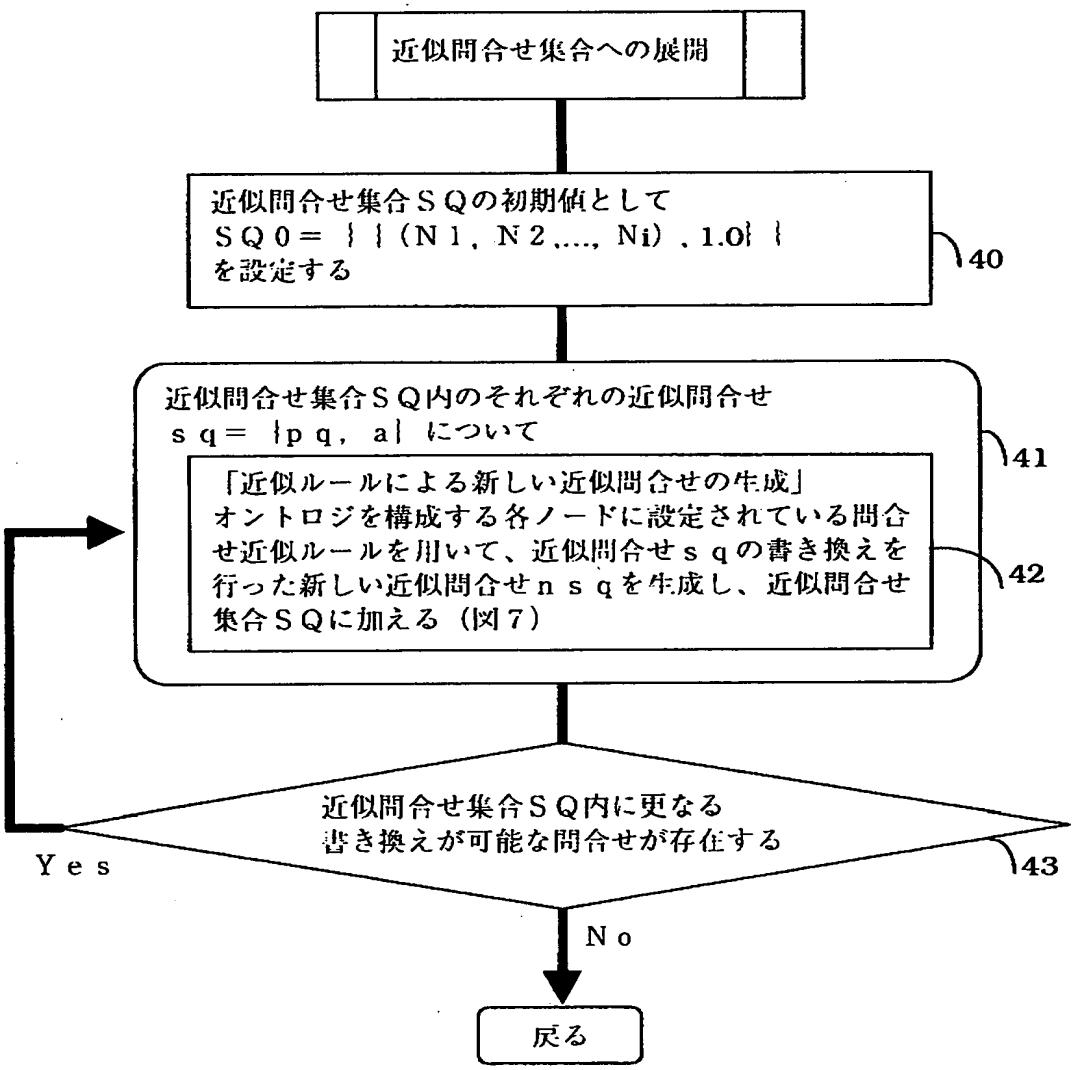
【図３】

図３


開始
問合せ文及び近似度の下限と実行コストの上限が指定された問合せを受付ける
30
問合せ内容をオントロジのノードと対応付ける
31
「近似問合せ集合への展開」
投入された問合せをオントロジのノード間の近似ルールに従って、指定された近似度を下回らない範囲で近似問合せの集合に変換する（図４）
32
「最適化近似問合せの選択」
各要素データベースや検索ツールの適用条件を参照して、近似問合せの内、実行可能なものを抽出する。ついで、抽出された近似問合せの集合から、投入された問合せを最も良く近似すると思われる近似問合せの組合せを抽出する（図６）
33
近似問合せ集合内の近似問合せを近似度の降順に整列
34
「問合せ実行プランの併合」
近似問合せの問合せ実行プランを近似度の大きいものから順に共通部分処理の括り出しなどを行いながら併合を行う。累計の実行コストが指定された上限を超えない範囲で問合せの併合を終了する（図７）
35
生成された実行プランに従って問合せ処理を実行する
36
終了

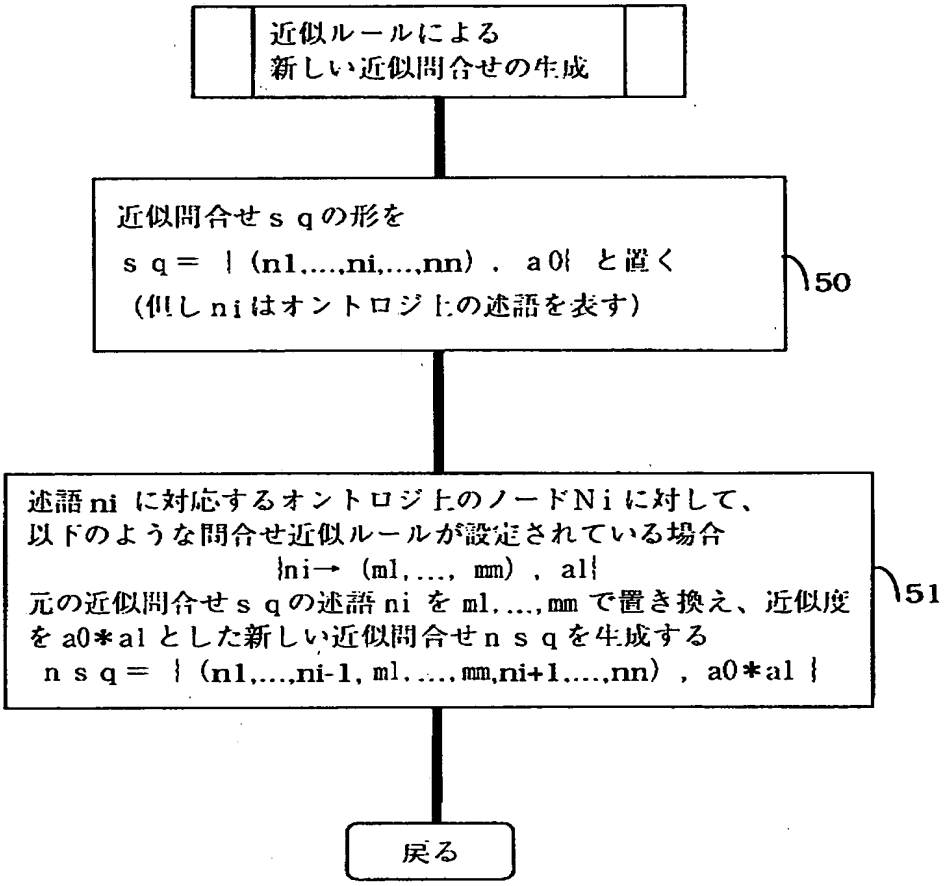
【図４】

図４


近似問合せ集合への展開
近似問合せ集合ＳＱの初期値として
ＳＱ０＝｛｛（Ｎ１，Ｎ２，…，Ｎｉ），1.0｝｝
を設定する
40
近似問合せ集合ＳＱ内のそれぞれの近似問合せ
ｓｑ＝｛ｐｑ，ａ｝について
41
「近似ルールによる新しい近似問合せの生成」
オントロジを構成する各ノードに設定されている問合せ近似ルールを用いて、近似問合せｓｑの書き換えを行った新しい近似問合せｎｓｑを生成し、近似問合せ集合ＳＱに加える（図７）
42
近似問合せ集合ＳＱ内に更なる
書き換えが可能な問合せが存在する
43
Ｙｅｓ
Ｎｏ
戻る

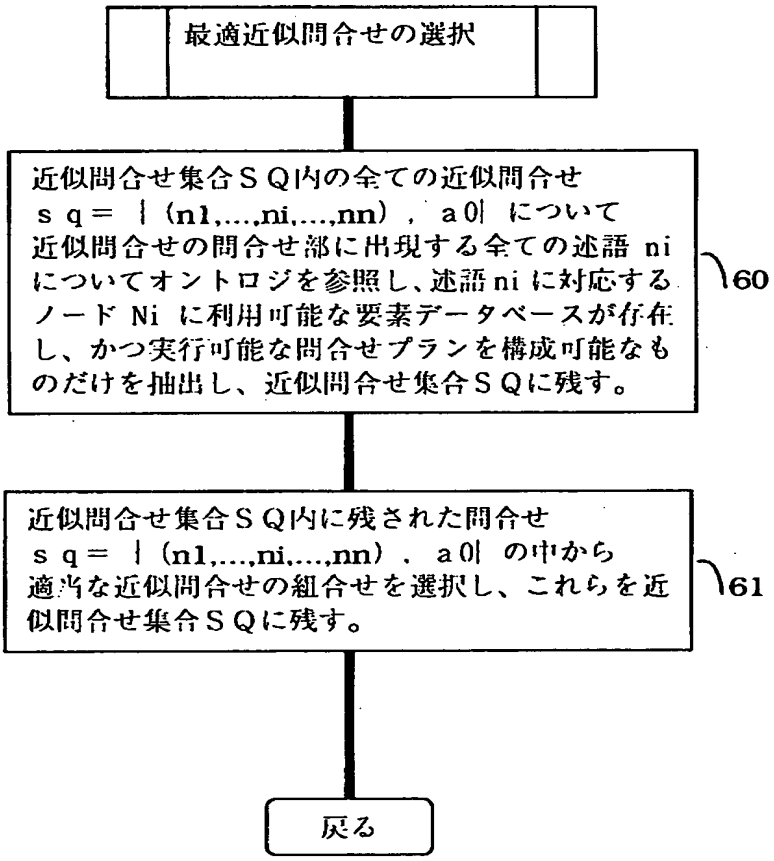
【図５】

図５


近似ルールによる
新しい近似問合せの生成
近似問合せｓｑの形を
ｓｑ＝｜(n1,...,ni,...,nn)，a0｜と置く
(但しniはオントロジ上の述語を表す)
50
述語ni に対応するオントロジ上のノードＮｉに対して、
以下のような問合せ近似ルールが設定されている場合
｜ni→ (m1,..., mm)，a1｜
元の近似問合せｓｑの述語ni を m1,...,mm で置き換え、近似度
をa0＊a1とした新しい近似問合せｎｓｑを生成する
ｎｓｑ＝｜(n1,...,ni-1, m1,...,mm,ni+1,...,nn)，a0＊a1｜
51
戻る

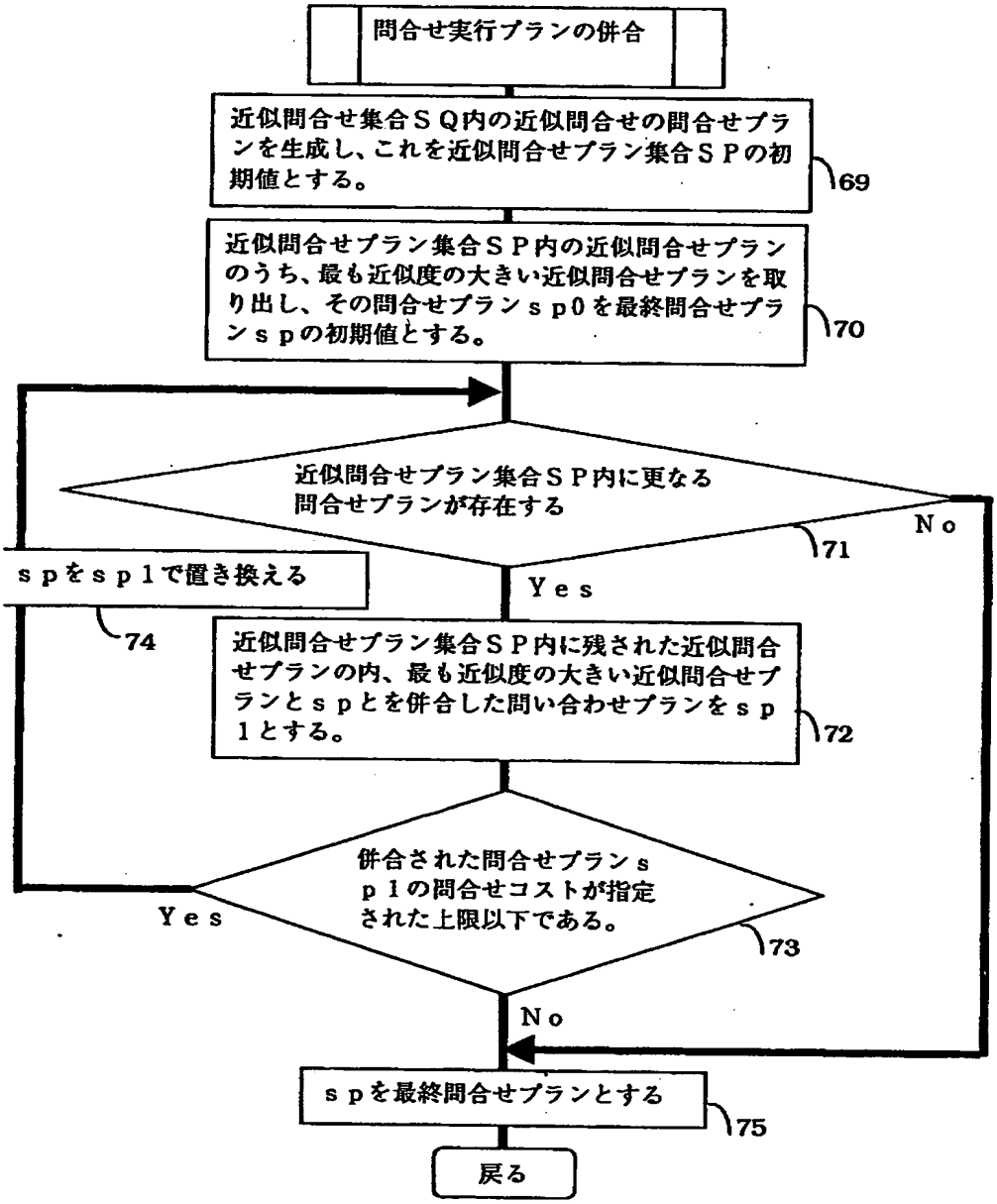
【図６】

図６


最適近似問合せの選択
近似問合せ集合ＳＱ内の全ての近似問合せ
ｓｑ＝｜(n1,...,ni,...,nn)，ａ0｜について
近似問合せの問合せ部に出現する全ての述語 ni
についてオントロジを参照し、述語 ni に対応する
ノード Ni に利用可能な要素データベースが存在
し、かつ実行可能な問合せプランを構成可能なも
のだけを抽出し、近似問合せ集合ＳＱに残す。
60
近似問合せ集合ＳＱ内に残された問合せ
ｓｑ＝｜(n1,...,ni,...,nn)，ａ0｜の中から
適当な近似問合せの組合せを選択し、これらを近
似問合せ集合ＳＱに残す。
61
戻る

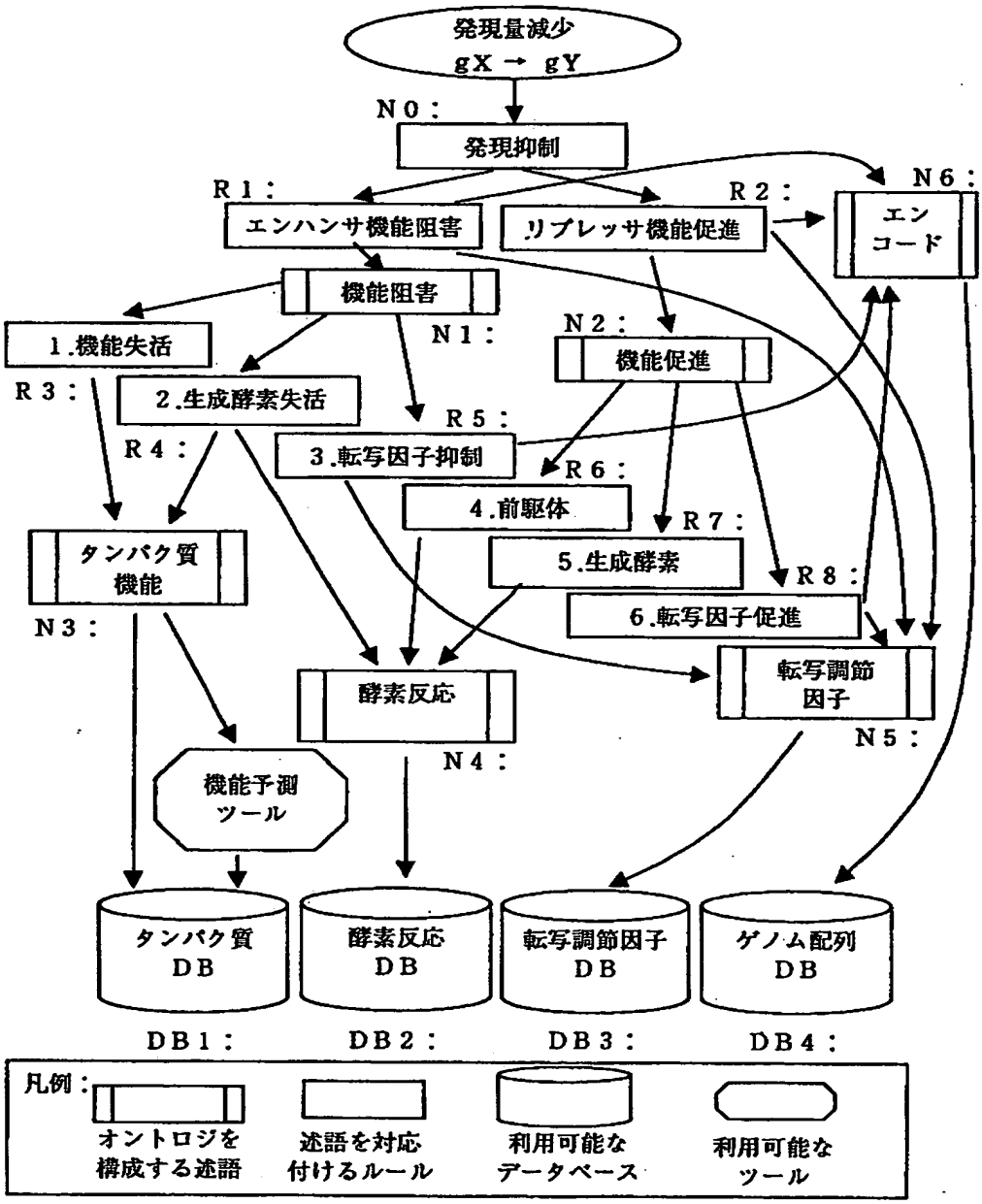
【図７】

図７


問合せ実行プランの併合
近似問合せ集合ＳＱ内の近似問合せの問合せプランを生成し、これを近似問合せプラン集合ＳＰの初期値とする。
69
近似問合せプラン集合ＳＰ内の近似問合せプランのうち、最も近似度の大きい近似問合せプランを取り出し、その問合せプランｓｐ0を最終問合せプランｓｐの初期値とする。
70
近似問合せプラン集合ＳＰ内に更なる問合せプランが存在する
71
No
Ｙｅｓ
ｓｐをｓｐ１で置き換える
74
近似問合せプラン集合ＳＰ内に残された近似問合せプランの内、最も近似度の大きい近似問合せプランとｓｐとを併合した問い合わせプランをｓｐ１とする。
72
併合された問合せプランｓｐ１の問合せコストが指定された上限以下である。
73
Ｙｅｓ
No
ｓｐを最終問合せプランとする
75
戻る

【図８】

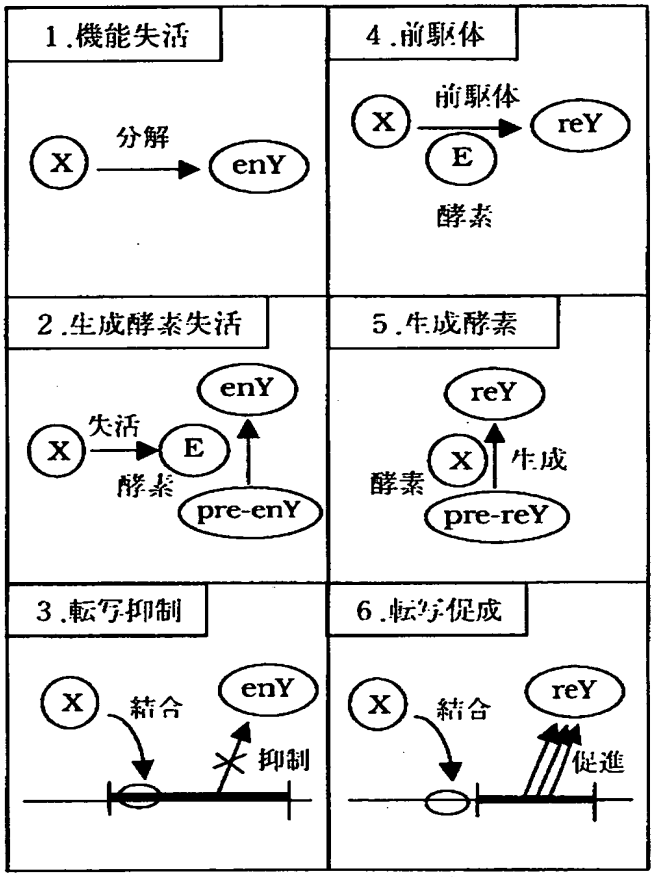
図８


発現量減少
gX → gY
N0：
発現抑制
R1：
R2：
N6：
エンハンサ機能阻害
リプレッサ機能促進
エン
コード
機能阻害
N1：
N2：
1.機能失活
機能促進
R3：
2.生成酵素失活
R5：
R4：
3.転写因子抑制
R6：
4.前駆体
R7：
タンパク質
機能
5.生成酵素
R8：
N3：
6.転写因子促進
転写調節
因子
酵素反応
N5：
N4：
機能予測
ツール
タンパク質
DB
酵素反応
DB
転写調節因子
DB
ゲノム配列
DB
DB1：
DB2：
DB3：
DB4：
凡例：
オントロジを
構成する述語
述語を対応
付けるルール
利用可能な
データベース
利用可能な
ツール

【図９】

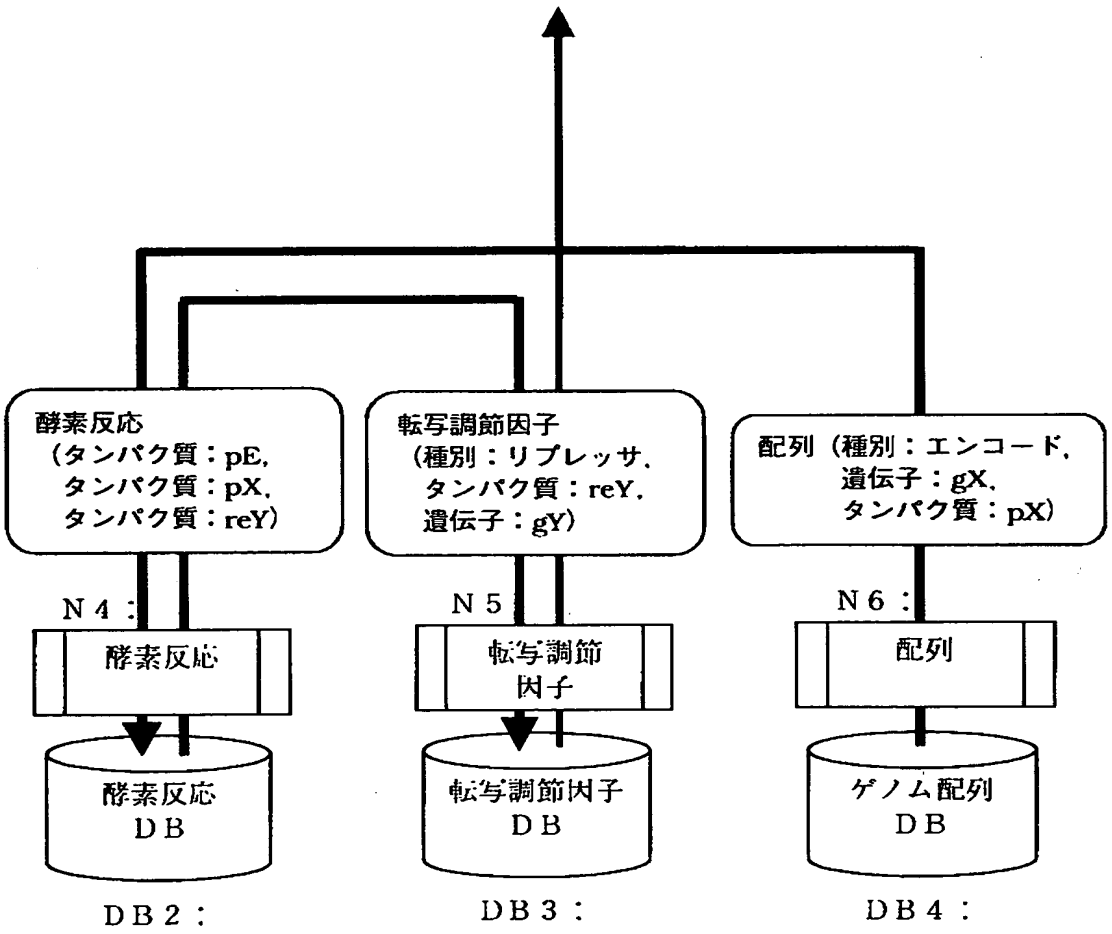
図９

近似ルールに対応する生体分子関連模式図：



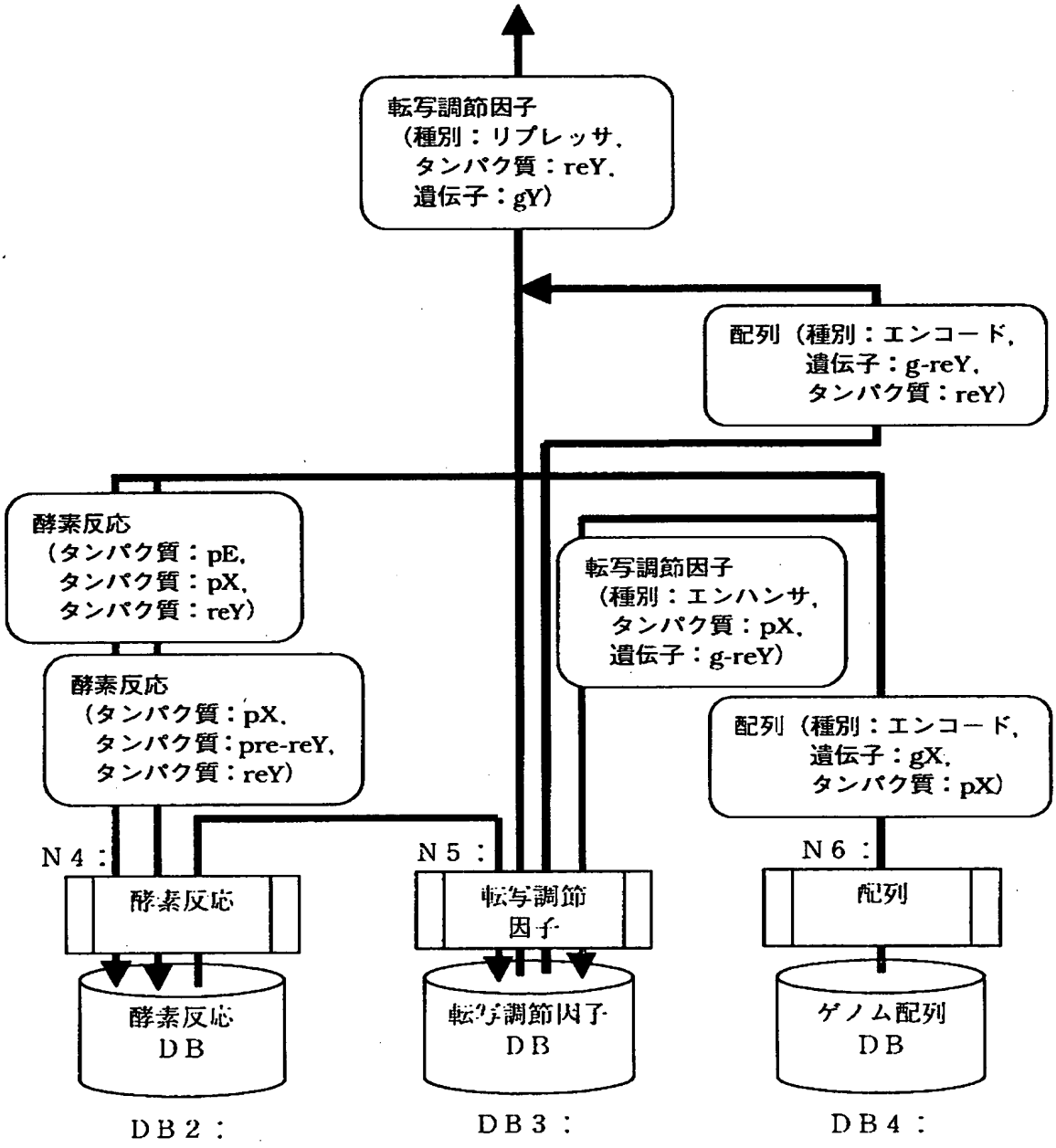
【図１０】

図１０


酵素反応
（タンパク質：pE，
タンパク質：pX，
タンパク質：reY）
転写調節因子
（種別：リプレッサ，
タンパク質：reY，
遺伝子：gY）
配列（種別：エンコード，
遺伝子：gX，
タンパク質：pX）
Ｎ４：
酵素反応
Ｎ５：
転写調節因子
Ｎ６：
配列
酵素反応ＤＢ
転写調節因子ＤＢ
ゲノム配列ＤＢ
ＤＢ２：
ＤＢ３：
ＤＢ４：

【図１１】

図１１


転写調節因子
（種別：リプレッサ，
タンパク質：reY，
遺伝子：gY）
配列（種別：エンコード，
遺伝子：g-reY，
タンパク質：reY）
酵素反応
（タンパク質：pE，
タンパク質：pX，
タンパク質：reY）
転写調節因子
（種別：エンハンサ，
タンパク質：pX，
遺伝子：g-reY）
酵素反応
（タンパク質：pX，
タンパク質：pre-reY，
タンパク質：reY）
配列（種別：エンコード，
遺伝子：gX，
タンパク質：pX）
N４：
酵素反応
N５：
転写調節
因子
N６：
配列
酵素反応
ＤＢ
転写調節因子
ＤＢ
ゲノム配列
ＤＢ
ＤＢ２：
ＤＢ３：
ＤＢ４：

【書類名】　要約書

【要約】

【課題】スキーマ構造や問合せ能力が異なる複数のデータベースを統合して、なるべく多くの問合せ結果を効率良く取得するための問合せプランを生成する問合せ最適化方法を提供する。

【解決手段】問合せに用いられる述語及び外部データベースを予め決められたオントロジーに従って関連付け、オントロジーを通じた問合せの発行を可能にする。このとき統合データベースに対して投入された問合せを、オントロジーが備える近似ルールに従って複数の近似問合せの集合に展開し、最適化モジュールでは外部データベースの問合せ能力を参照しながら、これらの複数の近似問合せが全体として効率良く行われるような問合せ最適化を行う。

【効果】統合データベースに対して発行された問合せを、外部データベースを組合せて処理する何通りもある問合せプランの中から一つを選んで実行する従来の問合せ最適化方式と比べ、外部データベースから取り出すことのできる問い合わせ結果の範囲を効率良く拡大することが出来る。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００１－０５３４７４
受付番号	５０１００２７９５１７
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１３年　３月　１日

＜認定情報・付加情報＞

【提出日】　平成13年 2月28日

次頁無

出　願　人　履　歴　情　報

識別番号　　〔０００００５１０８〕

１．変更年月日　　１９９０年　８月３１日
　［変更理由］　　新規登録
　　住　所　　東京都千代田区神田駿河台４丁目６番地
　　氏　名　　株式会社日立製作所